The Language Library: Many Layers, More Knowledge

Nicoletta Calzolari, Riccardo Del Gratta, Francesca Frontini, Irene Russo
Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche
Via Moruzzi 1 56126 Pisa, Italy
{name.surname}@ilc.cnr.it

Abstract

In this paper we outline the general concept of the Language Library, a new initiative that has the purpose of building a huge archive of structured colletion of linguistic information. The Language Library is conceived as a community built repository and as an environment that allows language specialists to share multidimensional and multi-level annotated/processed resources. The first steps towards its implementation are briefly sketched.

1 Introduction

In Natural Language Processing technologies even small amounts of annotated data can contribute to improve the performance of complex systems (Palmer and Xue, 2010). This evidence has led to the creation of many annotation schemes that encode our knowledge of syntactic, semantic and pragmatic features of every language.

Annotation is at the core of training and testing systems, i.e. at the core of NLP. Relations among phenomena at different linguistic levels are at the essence of language properties but we are currently over-simplifying annotation tasks, focusing mostly on one specific linguistic layer at a time, without (having the possibility of) paying attention to the relations among the different layers. At the same time our efforts are too much scattered and dispersed without much possibility of exploitation of others' achievements.

Today we have enough capability and resources for addressing the complexities hidden in multi-layer interrelations. Moreover, we can exploit today's trend towards sharing for initiating a collective movement that works towards creating synergies and harmonisation among different annotation efforts that are now dispersed.

In this paper we present the Language Library, an initiative which is conceived as a facility for gathering and making available through simple functionalities all the linguistic knowledge the field is able to produce, putting in place new ways of collaboration within the LRT community.

The rationale behind the Language Library initiative is that accumulation of massive amounts of (high-quality) multi-dimensional data about language is the key to foster advancement in our knowledge about language and its mechanisms, in particular for finding previously unnoticed interrelations between linguistic levels. The Language Library must be community built, with the entire LRT community providing data about language resources and annotated/encoded language data and freely using them.

With the Language Library we thus want also to enable/promote a more global approach to language studies and start a movement aimed at – and providing the facilities to – collecting all possible annotations at all possible levels.

Given the state of the art of linguistic annotation, we can certainly hope to gather tens of different annotation layers and types on the same data; once this is obtained, it will allow for a better analysis and exploitation of language phenomena that we tend to disregard today. In particular, interesting interrelations are likely to become visible among levels that are not often considered together, thus leading to improved computability (e.g. a coreference annotation on top of simpler annotation layers would improve machine translation performance). Part of this multi-layer and multi-language annotation should be performed on parallel (or at least comparable) texts, so as to foster comparability of new achievements and equality among languages.

Even if the Language Library will contain all kinds of processed linguistic data, in this paper we concentrate on the frequent case of annotated data.

Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 93–97,
Chiang Mai, Thailand, November 12, 2011.

2 Outline/General Concept

The Language Library is conceived as open and accessible repository where the language technology community can access and share corpora enriched with several layers of linguistic annotation. The Library is going to be:

- open, in that its content will be accessible to the community without restrictions;
- multilingual and multi-domain;
- multi-user and community oriented;
- multi-dimensional, containing multiple layers of annotation of the same text, possibly by multiple contributors;
- collaborative, in the sense of collaboration among experts, and also academics and NLP companies;
- reuse-oriented, promoting the reuse of annotated resources and annotation schemes;
- maintainable, endorsing the use of annotation standards;
- scalable, starting with a demo version with a limited number of texts and then progressively adding new features.

In order to reach this goal, a first population round of the Language Library will start around a core of parallel/comparable texts that will be annotated over and over again by several contributors submitting a paper for LREC2012. Hopefully the selected texts will be annotated with different tools and annotation schemes. The more this core grows, the more new contributors will be encouraged to participate by the possibility of building on existing layers of annotation to develop their own, which will be in turn added to the resource and become available to the NLP community. Notice that the Library should also be seen as a space where the theoretical and the applied linguistics communities could meet, in that the provided annotation can be both manually and automatically produced.

It is possible to envisage a scenario where an annotation layer (e.g. a human made annotation of coreferences on a portion of the texts from the Library) is first submitted by one author/researcher, used by another as a training set to tag a larger amount of the available texts, and then finally resubmitted enriched in size to be (at least partially) human checked again. By recursively doing so the Language Library could come to contain a great number of human checked sections, alongside increasingly accurate machine tagged ones.

In later stages the Library will grow both vertically by adding annotation layers and horizontally, by adding languages, domains, and by increasing the size of the corpora. At this point the possibility of comparing and cross-examining information from several annotation layers at a large scale will start to show its benefits both theoretically and in an NLP perspective.

In its mature stages the Library will consolidate by focusing on the enhancement of interoperability, by encouraging the use of common standards and schemes of annotation. It has to be underlined that the Language Library is conceived as a theory-neutral space which will allow for several annotation philosophies to coexist. The interoperability effort should not be seen as a superimposition of standards but rather as the promotion of a series of best practices that might help other contributors to better access and easily reuse the annotation layers provided.

In this sense encouraging the use of a representation format such as GrAF (Ide and Suderman, 2007) in a second stage might be helpful. On the one hand the stand-off approach of keeping each layer of annotation separated from the others and from the raw data seems particularly suitable for the Library; on the other hand GrAF enables a soft approach to interoperability, in that it can be used to uniformly represent formats that are both syntactically and semantically different and this could make it easier for the contributors to recognize compatible layers without forcing the adoption of a rigid standard. GraF converters from some known formats are currently under development and might be made available in the Language Library.

2.1 Building a community

As witnessed in the evolution of other collaborative resources, in order to attract the contribution of the community it is necessary to bring the Language Library to a level where the burden and the relative cost of sharing the resources is paid back by the possibility of accessing the resources released/produced by other researchers. In order to facilitate this the project will be built around existing frameworks of language resource sharing and

around their existing communities.

A number of ongoing initiatives (FLaReNet, META-SHARE and CLARIN among others) have already attracted around themselves a growing LRT community that requires consolidation of its foundations and steady increase of its major assets, minimising dispersion of efforts and enabling synergies based on common knowledge and collaborative initiatives.

The Language Library initiative will build upon the large experience gathered and the best practices and tools developed in these projects, both in terms of documentation and of collection and storage of the resources. While these initiatives have concentrated so far on language resources and tools, with the Language Library - started as a FLaReNet[1] initiative - focus will shift mostly on linguistic knowledge.

Most specifically the Language Library will be strictly connected with the following initiatives:

- The LRE Map (Calzolari et al., 2010), started at LREC 2010, collecting metadata about Language Resource and Technology;
- META-SHARE (Piperidis et al., 2011), an open platform providing an open, distributed, secure, and interoperable infrastructure for the Language Technology domain.

Both these initiatives rest on the assumption that availability is not enough: resources must be visible and easily retrievable. The Language Library will be made visible through META-SHARE, where a complete set of Metadata is already available for language resources and it can be immediately applied to describe and catalog the first nucleus of the Language Library.

2.2 Comparison with other initiatives

Recently other initiatives that share some points of similarity with the Language Library here described have been launched, proving the fact that the community is currently oriented towards similar goals.

The Manually Annotated Sub-Corpus (MASC)[2] of the American National Corpus (ANC)[3] is an open and downloadable corpus that that shares with the Language Library the idea of collecting as many annotation layers for a single text collection as possible. However, it is not conceived as a multilingual project and is more strictly limited to one corpus.

The Human Language Project (Abney and Bird, 2010) on the other hand is a multilingual project, that aims to build a Universal Corpus of the world's languages. In this case the immediate goal is to reach horizontal completeness (document as many languages as possible, with a special attention to endangered ones) and the project is specifically geared towards the Machine Translation community.

The Language Commons[4] finally is an online archive for the collection of written and spoken corpora in the open domain. The Language Library idea bears similarities to this experience, but it will dramatically shift the focus on the vertical dimension, in that it focuses also on gathering as many annotation levels for the same texts as possible.

3 First Experiment

After the success of the LRE Map[5] introduced for LREC 2010 and now used in many conferences as a normal step in the submission procedure (EMNLP and COLING among others), LREC 2012 will be the occasion to launch the LREC Language Library, that will constitute the first building block of the Language Library.

Because of the huge amount of data about resources provided for the LRE Map, we believe that times are ripe for the promotion of such collaborative enterprise of the LREC Community that will constitute a first step towards the creation of this very broad, community-built, open resource infrastructure.

Together with ELRA we will prepare as a first step an LREC Repository, part of the META-SHARE network, hosting a number of raw data on all modalities (speech, text, images, etc.) in as many languages as possible. When submitting a paper, authors will be invited to process selected texts, in the appropriate language(s), in one or more of the possible dimensions that their submission addresses (e.g. POS-tag the data, extract/annotate named entities, annotate temporal information, disambiguate word senses, transcribe audio, etc.) and put the processed data back in the

[1] www.flarenet.eu

[2] www.americannationalcorpus.org/MASC/Home.html (Ide et al., 2010)

[3] http://americannationalcorpus.org/

[4] www.archive.org/details/LanguageCommons

[5] www.resourcebook.eu

LREC Repository.

The processed data will be made available to all the LREC participants before the conference, to be compared and analyzed, and at LREC some/an event around them will be organized.

This collaborative work on annotation/transcription/extraction/... over the same data and on a large number of processing dimensions will set the ground for the future Language Library, linked to the LRE Map for the description of the data, where everyone can deposit/create processed data of any sort all our "knowledge" about language.

3.1 A case study: Annotation Resources at LREC2010

With the aim to highlight the feasibility of the LREC Repository for the Language Library, we propose a brief analysis of the annotation guidelines/tools inserted by authors as resources in the LREC2010 Map during the submission process. This will enable us to make, at this preliminary stage, an educated guess on the number and variability (with respect to languages, modalities, uses) of annotated texts that will be part of the core of the Language Library.

Amongst over 1990 resources, 62 are listed as "Representation-Annotation Formalism/Guidelines" ("R-A F/G" in Tables 1 and 2) while 136 are described as Annotation Tool ("AT" in Tables 1 and 2). Not every submission that report on the usage of an annotation tool provided also description for an annotation formalism, therefore its possible that more annotation schemes have been used.

As expected, the vast majority of annotation tools (see Table 1) are listed/described as language independent (82/136), while among RepresentationAnnotation Formalism/Guidelines 10/62 have been developed for English, 20/62 are language independent and 12/62 have been applied in multilingual resources.

	R-A F/G	AT
Language independent	20/62	82/136
English	10/62	18/136
Multilingual	12/62	11/136

Table 1: Most frequent values with respect to the language

Concerning modality, very few formalisms have been proposed for modalities other than Written (6/62), but among annotation tools 25/136 resulted useful for Multimodal/Multimedia modality, 8/136 for Sign Language and 7/136 for Speech modality.

The range of resource uses (see Table 2) is quite wide, with a prevalence of Knowledge Discovery/Representation (8/62, 6/136), Information Extraction, Information Retrieval (7/62, 13/136), Machine Translation, Speech To Speech Translation (6/62, 8/136). For Annotation Tool, Discourse, Acquisition and Dialogue are the other most frequent uses.

	R-A F/G	AT
Knowledge Discovery/Representation	8/62	6/136
Information Extraction, Information Retrieval	7/62	13/136
Machine Translation, Speech to Speech Translation	6/62	8/136

Table 2: Most frequent with respect to the uses

This information relative to a small subsets of the resources described by LREC2010 authors shows how in the starting phase the Language Library will be easily enriched with texts annotated on the basis of guidelines elaborated by scholars for a wide range of uses. Even if the incidence of languages other than English and of modalities other than Written is not so high, the existence of guidelines/formalisms focusing on more than one language represents an interesting chance to enrich the Language Library with more annotated data.

Finally, 40/62 Representation-Annotation Formalism/Guidelines have been listed as Newly created-in progress or Newly created-finished, a figure that shows how the Language Library can foster the knowledge about brand new annotation formalisms.

3.2 Future developments

In the initial phases of the project the main challenge will be to motivate the large parts of the community to join in the enterprise; subsequently more steps will be taken in order to enhance interoperability and avoid the proliferation of various, slightly different but incompatible annotation schemes.

In order to improve this the platform should make annotation schemes and tools available to the users, in such a way as to encourage the sharing and use of already existing standards. Ideally the platform could at some stage enable the hosting of on-line annotation tools, thus becoming a virtual environment for the recruitment of annotating workforce in a crowd-sourcing modality.

Also the dimensions and the modality of annotated data will have to be taken into account: we hope that not just small written corpora will be annotated and that the efficient management of audio and video files will be allowed in the platform.

4 Conclusions

It has been recognized that Natural Language Processing is a data-intensive discipline, so the LR community must now be coherent and take concrete actions leading to the coordinated gathering − in a shared effort − of as many (annotated-encoded) language data as it is able to produce.

In doing this a positive inspiration can be drawn from the success of similar experiences in other disciplines, e.g. astronomy/astrophysics, where the scientific communities cooperate in accumulate huge amounts of observation data for better understanding the universe. The most significant model is the recent successful effort for the mapping of human genome. The Language Library could be considered as a sort of big Genome project for languages, where the community will collectively deposit/create increasingly rich and mutilayered linguistic resources, enabling a deeper understanding of the complex relations between different annotation layers.

Acknowledgments

We thank the META-NET project (FP7-ICT-4 249119: T4ME-NET) for supporting this work.
The Language Library started as an initiative within FLaReNet - Fostering Language Resources Network (Grant Agreement No. ECP-2007-LANG-617001).
The Language Library has been discussed, among others, with Khalid Choukri, Thierry Declerck, Olivier Hamon, Joseph Mariani, Stelios Piperidis.

References

Steven Abney and Steven Bird. 2010. The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 88–97, Uppsala, Sweden, July. Association for Computational Linguistics.

Nicoletta Calzolari, Claudia Soria, Riccardo Del Gratta, Sara Goggi, Valeria Quochi, Irene Russo, Khalid Choukri, Joseph Mariani, and Stelios Piperidis. 2010. The LREC Map of Language Resources and Technologies. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Nancy Ide and Keith Suderman. 2007. GrAF: A Graph-based Format for Linguistic Annotations. In *Linguistic Annotation Workshop, ACL 2007*, pages 1–8, Prague.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. Masc: The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73. Association for Computational Linguistics.

Martha Palmer and Nianwen Xue. 2010. Linguistic annotation. In *The Handbook of Computational Linguistics and Natural Language Processing*, Blackwell Handbooks in Linguistics, pages 238–270. John Wiley & Sons.

Stelios Piperidis, Calzolari Nicoletta, and Maria Koutsombogera. 2011. META-SHARE: Design and Governance (Deliverable D6.2.1). Technical report, METANET, January. Dissemination Level: Restricted.